III.A.2.b. PC-Based Software

We began the lending library concept last year by making Kermit (a public domain terminal emulator and file transfer protocol from Columbia University) available to BIONET.We send them a diskette and documentation which they copied and returned to us. We currently make Kermit available for Apple II, Macintosh, IBM-PC and TRS-80 model computers. This year we have extended our lending library to include many BIONET-user developed programs and a number of utility programs that are useful for file transfer on IBM-PC and Macintosh computers. A complete list of the programs which are available is in Appendix IV. The <PC-SOFTWARE.MAC> and <PC-SOFTWARE.IBM> directories in particular have been referenced an average of 17 times a month each. In addition we have sent out diskettes 32 times that contain these utility programs. Their primary functions are to allow users to share software by helping them to upload and download files from BIONET and other mainframe computers.

We have also set up bulletin boards concerning PC-SOFTWARE in addition to PC-COMMUNICATIONS and these have been among the most popular bulletin boards with 107 and 47 message postings respectively last year. A number of bulletins give detailed procedures for performing electronic transfer of both text and binary files.

RNAFOLD AND BIOFLD - This year Dr. Zuker developed an IBM-PC version of his mainframe program. When this version is run on an IBM AT with 640 K of memory it is capable of calculating the minimum free energy secondary structure of RNAs' up to 360 bases long. His program has been downloaded from BIONET 51 times and in addition we have sent out copies of this program from the lending library.

MOLECULE - One of the weak aspects of Zuker's BIOFLD and RNAFOLD program is the display of the predicted secondary structure of the RNA. Both versions print a one dimensional representation of these inherently two dimensional structures. Dr. John Thompson (a post-doc in Professor John Woolford's group at Carnegie Mellon University) has written a program which displays the optimal secondary structure in two dimensions on a graphics screen. The MOLECULE program uses the connection output file (CT file) that is produced by both Zuker's RNAFOLD and BIOFLD programs. It may also utilize files produced by the NUCSHO program developed at NIH by Richard Feldemann. All these programs for displaying RNA secondary structure are available on BIONET. Appendix V shows a graphics display of a minimal energy secondary structure for a viroid RNA molecule as presented by the MOLECULE program. This program has been downloaded 51 times (average of 10 times per month) and in addition we have sent out copies of this program from the lending library.

ALIGN - Dr. Dan Davison has made an IBM-PC version of the program available for people to download and for our lending library. The IBM-PC version of ALIGN has been downloaded 20 times.

DM - Drs. Bruce Conrad and David Mount have developed a significant library of DNA analysis programs for use on IBM-PCs. They have made these programs available through a distributor who charges handling fees only. They also volunteered to place them in our public domain directories. BIONET users have downloaded this set of programs 31 times (about 5 times a month on the average). This is quite impressive usage considering the size of these programs (362K). These programs carry out editing, display and plotting of both DNA and protein sequences. They also have a complete restriction enzyme analysis package. They can plot dot matrix comparisons between sequences and have a very sophisticated oligonucleotide search capability. This is one of the most comprehensive software packages for sequence analysis in the public domain and we are very grateful to Dr. David Mount for making it available to the BIONET facility.

BIONET has always provided rapid updates to all the major collections of sequence data. These include the GenBank and EMBL nucleotide sequence collections and the NBRF/PIR Protein Data Bank. There are many other types of data of use to the community, however. Recently, we have been encouraged by Dr. Rich Roberts of our National Advisory Committee to view as legitimate Collaborative Research the collection and dissemination of various data sets related to molecular biology. The following summarizes our current and projected activities in this area:

- **Restriction Enzyme Database**. We continue to provide the community with the latest additions to the Restriction Enzyme Database, through the cooperation of BIONET and Dr. Roberts at Cold Spring Harbor (CSH). As described in our last Annual Report, modifications are uploaded automatically to BIONET shortly after they are incorporated into the on-line database at CSH. In addition, we provide the community with subsets of the list of enzymes that are commercially available.

- **VectorBank^{tm}**. IntelliGenetics has made substantial additions to its databank of map and sequence data for plasmids. This will shortly be released as part of updates to the Core Library, and thus will be available to BIONET. We have also urged the inclusion of the EMBL plasmid sequences in the EMBL database.We will add these to VectorBank as soon as they are available.

- **GNOMIC**. The publication *GNOMIC, A Dictionary of Genetic Codes*, by E.N. Trifonov and V. Brendel is an indexed compilation of important short nucleotide sequences. There are many entries in this compendium that may be used in database searches. We plan to enter these sequences, with annotations, into a format compatible with the *QUEST* program for pattern searching in biological databases. These will be made available as new "key" files which are used directly in *QUEST*.

- **New Sequences**. We will shortly establish a directory to which scientists can contribute new sequence data for examination by others prior to their formal acceptance and inclusion in the major databases. The <NEW-SEQUENCES> directory should be ready for user contributions around the end of December 1986 and a SEQUENCES bulletin board may be implemented to announce contributions. Sequences that have been accepted for publication will be differentiated from other contributions to the directory, and a simple index scheme will be provided for files in the directory. The current plan is to leave the responsibility for data quality control with the sequence contributor, and users will be notified to this effect. This

effort may be expanded depending upon user response and the availability of BIONET staff time.

III.A.2.c. Liaison with Other Resources

Several accounts have been established on BIONET to promote interaction with other related Resources. The following is a summary of current sites with which we can exchange information, together with a brief list of sites that will be accessible when the full implementation of ARPANET gateways is established:

- **Molecular Biology Computer Research Resource.** The MBCRR, at the Dana-Farber Cancer Institute at Harvard, shares information through mail delivery via the GENE account and via the bulletin board system. Most recently, these facilities were used heavily in support of the New Hampshire meeting on *Macromolecules, Genes and Computers* sponsored by the MBCRR. The MBCRR also supplies BIONET with a formatted version of the Protein Data Bank, organized along functional lines, and we have made this available to the BIONET community;

- **Molecular Biology Information Resource.** The MBIR at Baylor College of Medicine communicates with BIONET through Dr. Lawrence's account for electronic communication.

- **Protein Identification Resource.** Information exchange with the PIR at the National Biomedical Research Foundation is through the account of Winona Barker, the Principal Investigator;

- **GenBank.** Information exchange with GenBank is via accounts set up for communication with both Bolt Beranek and Newman and Los Alamos National Laboratory. The former is via Dr. Howard Bilofsky's account. The latter, LANL, is via accounts set up for Dr. Walter Goad, with sub-accounts for key personnel on his staff. This interaction is more intensive due to the importance of information exchange on new sequence data and through their participation in the GenBank bulletin board and related activities (see the next section on bulletin boards). A prototype of an automated sequence submission utility has been developed and tested at BIONET during the last year. This program will allow users to log onto BIONET and use its facilities to input a sequence for transmission to GENBANK. It provides the style of interaction familiar to BIONET users, and permits them to use any existing data file with which they have been working on BIONET. The goal of this facility is to make it easier and faster to submit sequences.

- **Other Resources.** Our impending connection to ARPANET will make communication with international resources substantially easier. All such resources have access to ARPANET via gateways to JANET in the United Kingdom, EARN in Europe, ACSNET in Australia, and BITNET equivalents in other countries. Thus, access will soon be simple to the European Molecular Biology Laboratory, the Molecular Biology Information Service in Australia, and the Imperial Cancer Research Fund in England. We already exchange information with the European Molecular Biology Laboratory through a temporary link routed through the SUMEX-AIM Resource at Stanford University.

III.A.2.d. Bulletin Boards

The following bulletin board topics are currently available on the system.

Bulletin Board Name	Description
ASK-BIONET	User queries and consultant responses
BION	Information for BION workstation users
BIONET-NEWS	General BIONET announcements
CONTRIBUTED-SOFTWARE	Information on programs contributed by users
EMPLOYMENT	Job openings
GENBANK	New bulletin board in preparation for GenBank information and queries
GENE-EXPRESSION	Scientific interest group
GENOMIC-ORGANIZATION	Scientific interest group
INFO-1100	Computer interest group
INFO-AMIGA	Computer interest group
INFO-ATARI16	Computer interest group
INFO-IBM-PC	Computer interest group
INFO-KCC	Computer interest group
INFO-KERMIT	Computer interest group
INFO-LAW	Assorted legal information
INFO-MAC	Computer interest group
INFO-SUN-SPOTS	Computer interest group
INFO-VAX	Computer interest group
METHODS-AND-REAGENTS	For reagent exchanges and announcements about lab methods
MOLECULAR-EVOLUTION	Scientific interest group
ONCOGENES	Scientific interest group
PC-COMMUNICATIONS	Information on communications software
PC-SOFTWARE	General PC software announcements
PLANT-MOLECULAR-BIOLOGY	Scientific interest group
POLITICS	Nonpartisan activities
PROTEIN-ANALYSIS	Scientific interest group
YEAST-GENETICS	Scientific interest group

The leaders of the individual boards are:

ASK-BIONET	Dr. David Kristofferson
BION	Jaya Carl
BIONET-NEWS	Dr. David Kristofferson
CONTRIBUTED-SOFTWARE	(to be named)
EMPLOYMENT	Mary Warner
GENBANK	Dr. Christian Burks
GENE-EXPRESSION	Dr. William Sofer
GENOMIC ORGANIZATION	Drs. Robert Jones & Stephen Harris
METHODS-AND-REAGENTS	Dr. Larry Kedes
MOLECULAR-EVOLUTION	Dr. Dan Davison
ONCOGENES	Dr. David Steffen
PC-COMMUNICATIONS	Mary Warner
PC-SOFTWARE	Dr. Doug Brutlag
PLANT-MOLECULAR-BIOLOGY	Dr. Robert Jones
POLITICS	Dr. Michele Cimbala
PROTEIN-ANALYSIS	Amos Bairoch
YEAST-GENETICS	Dr. John Thompson

Note: All the INFO- bulletin board material is received from information sources outside of BIONET.

Several changes were implemented in the bulletin board system during the final months of 1986. Four older bulletin boards (IMMUNOLOGY, LIBRARIES, VECTORS, and LAB-METHODS) were combined into a METHODS-AND-REAGENTS board. This increases the total number of subscribers who see each message and as a result should facilitate exchange of information and materials. Over 130 investigators on the system now subscribe to this new bulletin board.

A new ASK-BIONET bulletin board was started for the purpose of posting user questions and consultant answers that are of interest to all BIONET users.

All of the previous bulletin boards were scrutinized for use and a few inactive boards were deleted. Enthusiastic new leaders were found for others and some new topics were introduced. Although BIONET had a policy of waiving the $400 subscription fee for bulletin board leaders, this policy was not previously advertised. Its announcement on the system stimulated additional inquiries. It is still too early to assess the results of all the recent changes, but the situation is being monitored carefully. A brief description of some of the new activity follows.

The ONCOGENES board has started anew under the leadership of Dr. David Steffen who is maintaining an up-to-date compendium of oncogenes on it. Dr. John Thompson has recently begun directing the YEAST-GENETICS board.

A new GENBANK bulletin board is being implemented. The board will be directed by Dr. Christian Burks at GenBank in Los Alamos and will serve to facilitate communication between BIONET users and GenBank.

Dr. Robert Jones is the new leader for the PLANT-MOLECULAR-BIOLOGY board. He also shares the leadership of the new GENOMIC-ORGANIZATION board with Dr. Stephen Harris.

Amos Bairoch, the developer of PC/GENE, is leading a new bulletin board on PROTEIN-ANALYSIS which will deal mainly with protein sequence analysis methods and protein sequence data banks, but is open for input on any matter concerning proteins.

III.A.3. Core Research

III.A.3.a. Multiple Sequence Alignment

We have recently started work on a large scale project to review all the presently developed multiple-sequence alignment software with the goal of making available to the BIONET community as many of these programs as possible. After having surveyed the community that is involved in this kind of software research and development, we approached all the major developers and asked both for detailed information about their programs and also whether they would be willing to donate them for possible use by the BIONET community. So far we have received programs from the following researchers: Bill Bains from the University of Bath, England; Osamu Gotoh from the Saitama Cancer Center Research Institute in Japan, Mark Johnson from the University of Southern California, Hugo Martinez from the University of California at San Francisco; and Joel Sussman from the Weizmann Institute in Israel. In addition, we are waiting for programs from Michael Waterman from the University of Southern California and Wayne Anderson from the University of Alberta, Canada. As we receive the above software, we will conduct in-depth testing of their performances and when we feel that the software would be useful to the BIONET community, will release it after any necessary revisions. We currently have running on the 2060 the first three of these programs: Bill Bain's MULTAN, Hugo Martinez's MALIGN, and Joel Sussman's PROT3.

Bill Bains' MULTAN program, (see W. Bains, NAR 14: 159-177), has been updated and now includes, among other things, a more generalized input file format, user-specified output files, and reformatted output. We released this updated version, XMULTAN, to the BIONET community at the end of September 1986 and approximately a month and a half after its release, there have been 233 runs of the program with a usage rate of 172 per month. This usage rate falls in the range of the rates for the IG library programs of between 89 and 2020 runs per month and is higher than for three of the IG programs: SIZER, MAP and CLONER. The feedback so far has been that XMULTAN is meeting the needs of the BIONET Community for a program that performs multiple sequence alignments and generates a consensus sequence. XMULTAN takes many DNA sequences (it easily handles 30) of up to 1500 base pairs in length and generates the alignment and subsequent consensus sequence. The algorithm for XMULTAN is heuristic, and the actual running time for sequences with a fairly high degree of similarity (approximately 75%) is reasonably short. They range from 1-2 minutes CPU for a small number of highly similar short sequences to 10 minutes CPU for a larger number of less similar long sequences. The consensus generation for 10 sequences, 400 base pairs long, and 82% similarity took 7.6 CPU minutes. Bill Bains is presently rewriting MULTAN to include improvements to the algorithm and we hope to provide his new version of MULTAN to the BIONET Community as well.

Joel Sussman's PROT3 program to align three protein sequences is based on an extension of the Needleman-Wunsch algorithm (see M. Murata, J.S. Richardson, and J.L. Sussman, PNAS 82:3073-3077,1986). The program requires a large amount of storage space to run and because of the

machine constraints of our DEC 2060 only three sequences of no larger than 60 amino acids will run. We are presently looking into ways around this situation. PROT3 was written for the VMS operating system and does not present a storage problem on VAX systems. Additionally, the performance of this program and the other programs that we have received (especially Gotoh's and Johnson's) need to be assessed before we make any final decisions.

III.A.3.b. BIONET Satellite Program

This program has the goal of distributing the BIONET Resource among computers throughout the academic community. At the same time we want to establish better communication links among BIONET, its Satellites, and other computing resources in molecular biology. We currently have Satellites established at the Salk Institute, at the US Department of Agriculture, and at Fort Dietrick (US Army RIID).

We are following two approaches to communication with other facilities-ARPANET and a dial phone line-based network that we are simply calling the BIONET Network for the moment. Significant development occurred on the the BIONET Network this year. We have installed software on the BIONET central resource that can communicate with other computers through the same terminal connections that serve scientists using the resource directly. Since the majority of the computers at current or potential Satellite sites are Digital Equipment Corporation VAXes running the VMS operating system, we have implemented communication software for this operating system. To augment the basic mail capability of VMS, we have also arranged for distribution of an improved mail delivery program to BIONET Satellites. The software is scheduled to be delivered to the first BIONET Satellite for testing in December 1986.

III.A.3.c. Hardware Text Searching Machines

In our last Annual Report we discussed our investigations of different machines designed for high speed searching of text for patterns of strings. We identified one machine, the Fast Data Finder made by TRW, as the clear leader for applications to pattern searching in biological sequences. This machine also has the potential for performing at least some aspects of the problem of determining sequence homologies. Over the past year, we have been negotiating with TRW to attempt to find some mechanism for BIONET's access to an FDF machine, but to no avail. We have been unable to meet their charges for access within our limited budget, although negotiations are continuing for access to less expensive versions of the FDF. Despite these problems, we continue to feel that the FDF is an extremely promising architecture, and we will do everything we can to arrive at an agreement with TRW.

In a parallel effort, we have entered into discussion with Dr. Peter Denning, head of the RIACS (Research Institute for Advanced Computer Science) project, headquartered at NASA/Ames Research Center. He

has a proposal submitted for purchase of a Connection Machine from Thinking Machines Corporation in Cambridge, MA. This is a processor that includes up to 64K individual processors operating in parallel. We have observed this machine in operation, performing high speed text searches, and its performance is impressive. BIONET has agreed with Dr. Denning to lead a Biotechnology Working Group, one of several groups that will explore a wide variety of applications of this novel architecture. At the time of writing of this report, final word on funding of the machine had not yet been received.

III.A.4. BIONET Training Program

Our training program during 1986-1987, emphasized the holding of regional trainings in different areas of the United States. The goal has been to try to draw BIONET users to trainings who would probably not otherwise attend. In addition to the shorter trainings at both the ASBC meeting in Washington, DC and the Miami Mid-Winter Symposia in Florida, we have held two magnet trainings during the summer of 1986: one in the northeastern United States at Dartmouth College in New Hampshire and one in the western United States at Stanford University in California. (See Appendix VI for an example of the mailer for BIONET training sessions). These trainings accomplished the main objective of training relatively novice users to more effectively utilize the system. Since these magnet trainings were successful, we plan to continue holding them in 1987.

In addition to the regional trainings we have also initiated a telephone training program for new users. The aim is to reach new users before they start using the system and assist them in getting started. Although we only began this program in August, the feedback that we have received so far indicates that it is worthwhile, and we plan to continue it.

The following summarizes the past year's training activites and those that are planned prior to the end of the current grant year. Training session schedules for all the trainings are listed in Appendix VII.

MIAMI MID-WINTER SYMPOSIA 1986 Training. BIONET sponsored a six hour training split between the afternoons of February 5-6, 1986 at the Miami Mid-Winter Symposia in Miami, Florida.

AMERICAN SOCIETY FOR BIOLOGICAL CHEMISTRY MEETING Training. We held a three hour class June 12, 1986 at the ASBC meeting in Washington, DC. The class was very well-attended (25 people) and we plan to hold future trainings at upcoming ASBC meetings.

DARTMOUTH Training. With much help from Bob Gross and Jo Steele at Dartmouth, BIONET held a two day training session on August 7th and 8th, 1986. There were 17 attendees, almost a third of whom (5) were from outside the area. The training was divided into novice (first day) and advanced (second day) with essentially all trainees attending both sessions. Dartmouth facilities included a VAX computer and terminal room for hands-on sessions for both days. The training was very well received.

STANFORD Training. We held a three day training for 43 people at the Stanford Business School's computer facility August 27-29. The three days were divided between novice, intermediate, and advanced topics. Although the training was rated highly by the attendees (80%), some felt that the number of trainees was too high. In the future we will try to limit class size to a more manageable 20-25 people so that more personal help can be given to the individual trainees.

MIAMI MID-WINTER SYMPOSIA 1987 Training. We plan to hold a training session on the evening of February 11, 1987 and it will be very similar to the previous year's session.

III.A.5. Resource Facilities

Previous reports have discussed the DEC-2060 and the various software and database libraries provided by the BIONET Resource. In this section we highlight significant changes and additions to the suite of hardware and software that comprise BIONET.

III.A.5.a. Computer and Telecommunication Networks

Hardware. The BIONET Central Resource Machine is a Digital Equipment Corporation 2060 computer. An Ethernet interface was added this year to provide access to ARPANET and other IntelliGenetics resources. The old DECNET front end, the DN20, was retired upon the introduction of the new Ethernet interface, the NI20. The MCA25 cache memory was ordered and scheduled for installation in December 1986, replacing the MCA20. This upgrade increases the throughput for the users of the resource.

The hardware configuration is as follows:

```
KL10-E Model R Processor:

      2 MF20/MG20 Memory controllers
   2 MW MG20 Memory
 .75 MW MF20 Memory
        MCA25 Cache Buffer Memory
      2 RH20 Massbus Channels
        NI20 Ethernet Interface

Console and Front End Processor:

        PDP-11/40 CPU, 32 KW 16 bit memory
        RX02 Dual floppy disk drives
      8 DH11 Terminal interfaces        8 * 16 TTY lines each = 128 lines
        RH11 Massbus Channel
        LP20 Line printer interface

 Peripherals:

      3 RP07 disk drives                111MW each
        RP06 disk drive                 39MW
                372 MW Total disk storage
```

```
TU78 1600/6250-BPI tape drive
LP26 600 LPM Line printer
Imagen Imprint-8/300 Laser Printer
```

Disk space (data storage)

Public structure (PS:) disk space use on the 2060 is dynamic. The following snapshot is representative of typical usage, and is taken from December 1986.

Total disk space	433,000	(pages--222 million words)
Overhead/Common	<143,000>	(Core, System and System Support Libraries)
Swapping Space	< 25,000>	
File system Overhead	< 67,000>	(Directories and index pages)
	198,000	
BIONET Allocation	99,000	(Half of the available space)
BIONET Usage 12/86	< 85,000>	
Unused space	14,000	(Available for BIONET growth)

Note that file system overhead varies greatly depending on the size of the files involved. Since BIONET users have many small files, BIONET growth may increase file system overhead, altering the above distribution.

Public Data Network Connection. BIONET is accessed principally over the Telenet Public Data Network, operated by US Sprint. An X.25 PAD (packet assembler/disassembler) is located on-site. This is known as the Host PAD, or HPAD. It provides individual terminal ports which are cross-connected to those on the DEC-20. The Telenet trunk line operates at 9600 baud synchronously, and the PAD converts this into up to 16 asynchronous ports whose speed is typically 1200 baud. A handshaking protocol is employed to smooth over bursts of data during the multiplexing.

Our former Public Data Network, UNINET, merged into Telenet in October 1986. BIONET had formerly made use of Telenet. UNINET was originally chosen as a replacement for Telenet because of its better response time and its lower cost. The lower cost was achieved through a very favorable fixed price per port arrangement that we negotiated with UNINET. The favorable pricing arrangement was renewed for a one year term beginning in July 1986, and continues to be honored by the new US Sprint Telenet.

During the year we increased from 12 to 16 the number of data network host ports used by BIONET, and usage is monitored carefully in the event more are needed. The ports are accessed in sequence, with those higher in the sequence not being used while any lower port is free. The number of connect hours per month drops off after the first 8 ports. The usage on these first 8 ports therefore represents many more

sessions than does the usage of ports 9 through 16. Our monitoring of the port use also has revealed that it would be cheaper for BIONET to lease the higher-numbered ports on a use, or traffic, basis. We currently are leasing 9 ports fixed, 7 on traffic, and will change this distribution as required for the lowest possible cost.

We had been examining the replacement of the leased HPAD supplied by UNINET with a BIONET owned HPAD. The consideration is the savings of lease charges while maintaining adequate reliability. We were unable to procure a suitable HPAD for BIONET's use on Uninet. Following the Uninet merger into Telenet, we are reexamining the issue, and we may purchase an HPAD from Telenet rather than leasing it.

ARPANET - During the course of the year, BIONET participated in two task force meetings to look at issues related to computer networking in the scientific community. This task force operates in conjunction with the DARPA Internet Activities Board. BIONET has arranged Internet access to ARPANET through a DARPA-funded project with IntelliCorp. In exchange for our assistance with the mechanics of the connection to ARPANET, BIONET will be able to make use of this connection for communications, especially electronic mail. The data connection to ARPANET took longer than expected, arriving in November 1986. The activation of the gateway computer at IntelliCorp is expected to take place during December. Since there are mail gateways from the ARPANET to other communications networks, this connection will do much to expand BIONET's reach, linking it with networks such as BITNET, EARN and CSNET as well as the DoD Internet.

Until the regular Internet connection is operational, BIONET has arranged mail access to the Internet with the valuable cooperation of the SUMEX-AIM NIH resource operated by Stanford. The same software that will link us to BIONET satellites allows us to exchange electronic mail through the SUMEX-AIM system. This has allowed scientists on BIONET to participate in information exchange using electronic mailing lists of Internet, which we import and make available as electronic bulletin boards on BIONET.

During the year, BIONET's central DEC-2060 resource was upgraded with the capabilities necessary for the Internet connection. An Ethernet interface now connects the DEC-20 to an IntelliGenetics local area network. An IntelliGenetics gateway connects that network to one at IntelliCorp, on which there will exist the gateway to ARPANET. Software for the TCP/IP protocols has been licensed and installed on the 2060.

III.A.5.b. Summary Statistics on Machine Use

The cpu cycles of the DEC-2060 computer are allocated to the user community, including BIONET, by the system's class scheduler. This scheduler is given the percentage of the machine to allocate to each class of users. Any cycles not consumed by a given class ("windfall")are available to the rest of the user community. This method was chosen so that cpu cycles not consumed by one segment of the community could be used by other segments if needed, i.e., no cpu cycles are wasted if someone needs them.

The current percentage allocations ("pieslices") are shown in Figure III-1. As summarized in the figure, BIONET Class I (and III and IV) are allocated 29% of the machine, and Class II and staff 9%. The 24% overhead (system overhead, batch and computer staff and operations) is allocated one-half to BIONET, for a total of 50%. The only major change to these allocations since last year was an increase from 20% to 24% in the amount allocated to system overhead, reflecting effects of a new release of the operating system. Because the operating system itself tends to consume less overhead, the effects of the increase in other system overhead on the users are minimized.

Note that the BATCH class is assigned 1% of the system during prime time. In non-prime time, the percentage allocation is increased substantially in response to demands by the BIONET community.

The actual use of the machine by the BIONET community has been substantially greater than 50% of the total cpu cycles allocated. As an example, the percentage use of the machine for the month of October, 1986 is shown in Figure III-2. It is clear that BIONET is receiving far more than its allocated share of the cpu cycles. Note that BIONET scientists' use of BATCH is charged to the individual accounts by the accounting program. Thus, extensive use of BATCH shows up in this pie chart as BIONET Class I (or II) use, rather than in the category BATCH Jobs.

The data for BIONET percentage of system use are plotted in histogram form in Figure III-3. This figure demonstrates that BIONET has utilized well over 50% of the total cpu cycles used on the 2060, and routinely consumes two-thirds of the total cpu cycles used on the system.

In the following series of tables and figures, we provide further details on the actual use of the system by the BIONET community. Looking first at use of the system in prime time (8 AM - 8 PM, M-F, PST), data for cpu time and connect hours for the indicated segments of the community are given in Tables III-3 and III-4 by month, and totals. The cpu data in Table III-3 is also plotted in histogram form in Figure III-4.

The main conclusion derived from these data is that the BIONET resource is being fully utilized. The data for this year shows only a slight upwards trend and is on the average about 25% higher than the corresponding cpu usage for the previous year.

Figure III-1: Pieslice Allocations of the DEC-2060 Computer

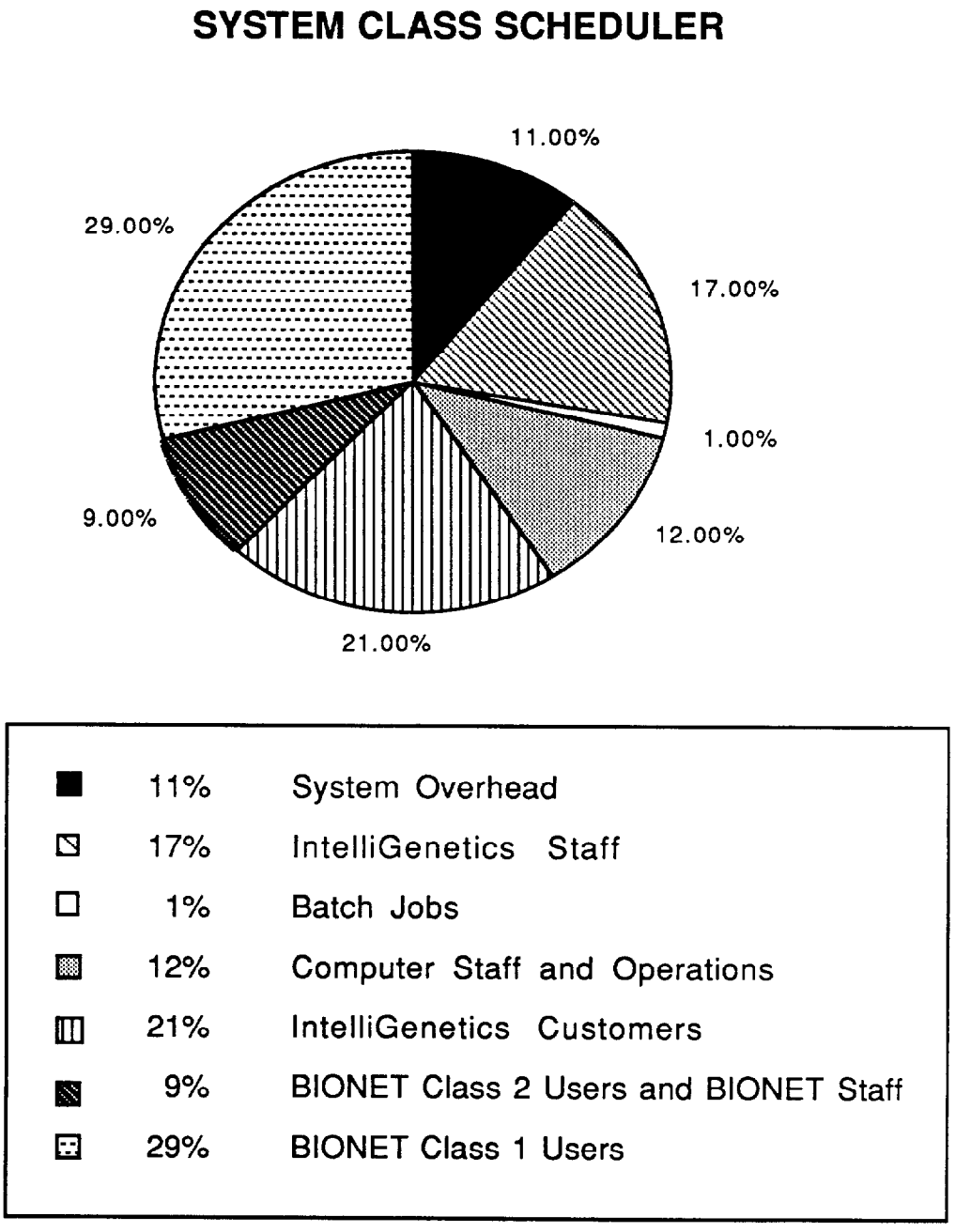


FIGURE III-2: Actual use of the DEC-2060 for the Month of October, 1986

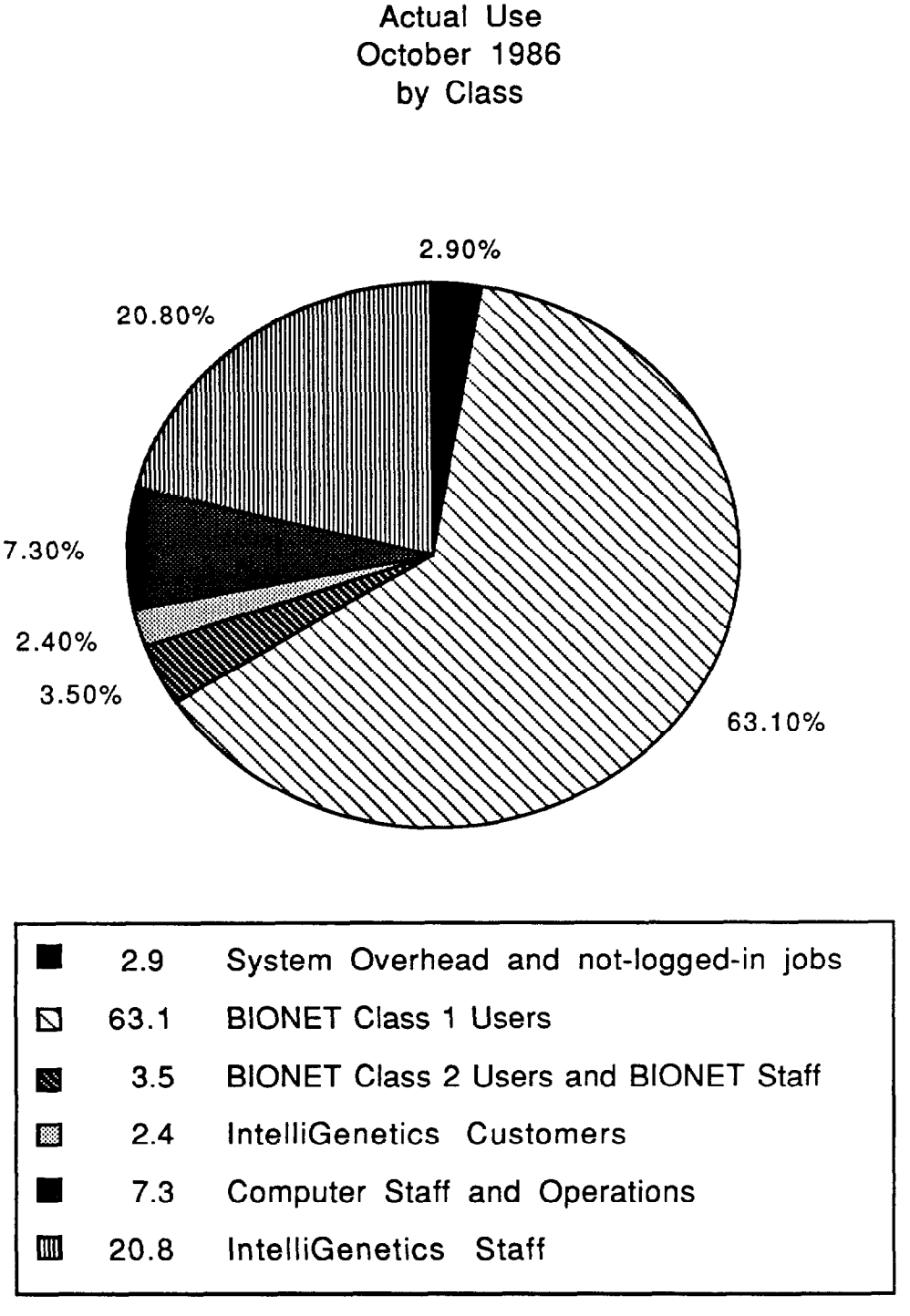
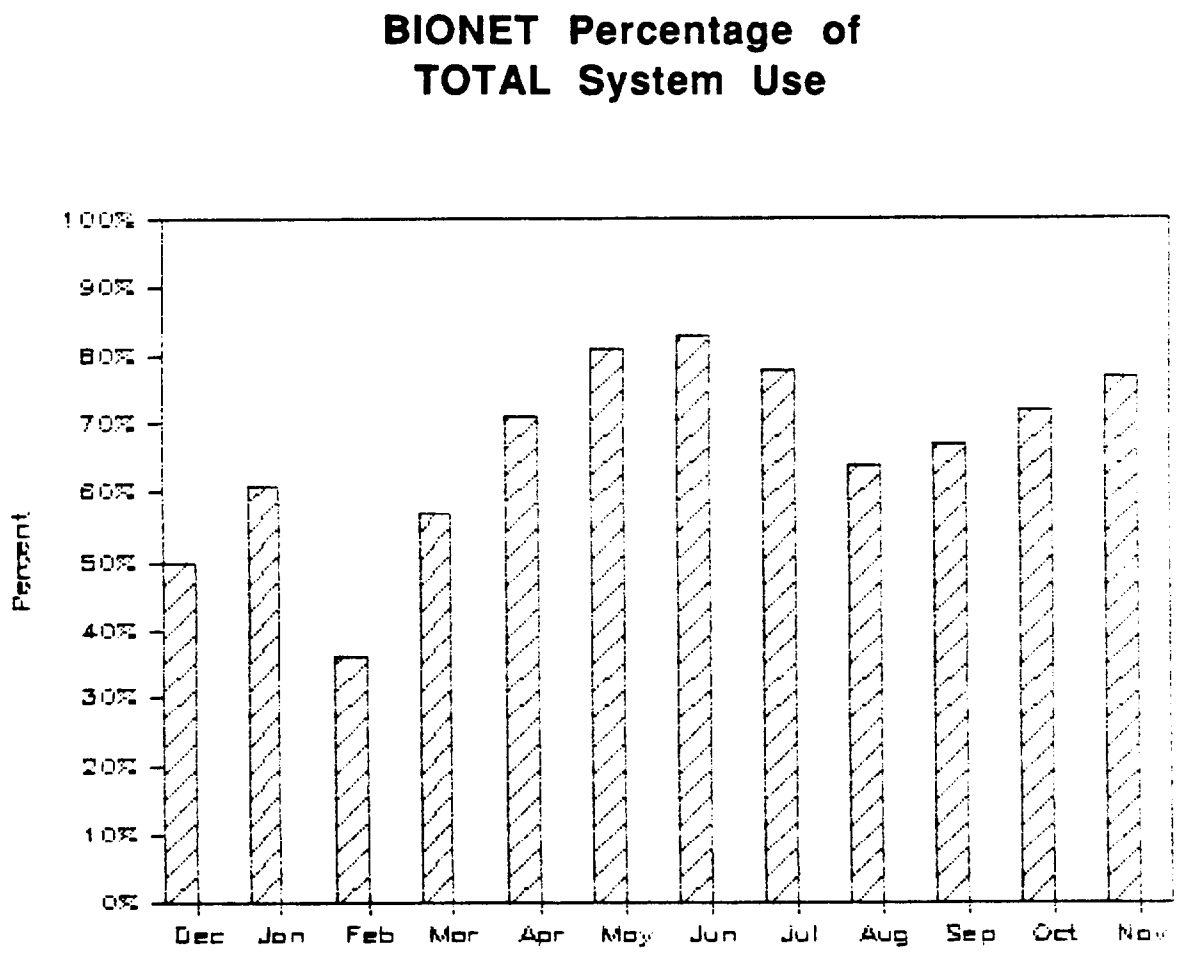


Figure III-3: BIONET's Percentage Use of the DEC-2060, 12/85 - 11/86



The total number of prime-time connect hours, (Table III-4), for BIONET Users is also up over last year by about 43%. The higher increase in connect time versus cpu time is probably a consequence of the increased demands on the system.

The data for non-prime time (weekends and 8 PM - 8 AM M-F) are shown in Tables III-5 and III-6, and the data on cpu time are plotted in histogram form in Figure III-5. Non-prime time cpu usage by BIONET has increased by around 75% over the course of the year. These increases are due primarily to the extensive use of overnight batch runs to perform time-consuming analyses involving database searches, using the IFIND homology and the QUEST database search and retrieval programs. Thus, the community has gravitated naturally toward off-hours use of these programs for such analyses. Given low use of the system by other classes in non-prime time, BIONET consumes most of the cpu cycles actually used during these times.

The data for total use of the Resource by BIONET are presented in Tables III-7 and III-8 and the total cpu time is summarized in Figure III-6.

One important conclusion from all these data is that the Resource is close to saturation. Certainly, during prime time, the system load is becoming a barrier to rapid computation. At this point, limitations on the number of access ports keep the load average under control by limiting the number of concurrent users.

Summary data for use of our telecommunications network are presented in Figure III-7 by month for the past 12 months' use of the UNINET (through the end of October, 1986) and Telenet (beginning October, 1986) networks. UNINET was taken over by Telenet during the course of this year, necessitating the change of networks.

III.A.5.c. Computer Software - Core Library

There have been two major releases of the IntelliGenetics software systems that make up the Core Software Library. This software is made available to the BIONET community immediately upon its formal release. The first release, in mid-1986, included major upgrades to the heavily used nucleic acid (*SEQ*) and protein (*PEP*) analysis programs, and the sequence homology program *IFIND*, plus many improvements to other modules. For example, a DIGEST command was provided in *PEP* to simulate digestion of a protein with a selectable set of chemical and biochemical reagents, calculate fragment sizes, and predict gel behavior of the fragments.

The second release, around the end of 1986, includes substantial changes to all programs except *MAP* and *SIZER*. Major improvements to the *GEL* program for constructing a consensus sequence from individual sequences have been made, including the ability to combine sequencing projects, and to produce laboratory notebook records of all steps of the gel sequence assembly. In addition, an extensive graphical

Table III-3: BIONET Prime Time CPU Minutes

	BIONET Users except staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	5732.4	409.5	360.6	6502.5
Jan	5293.4	425.4	337.7	6056.5
Feb	4406.8	234.8	268.4	4910.0
Mar	6181.9	384.2	346.8	6912.9
Apr	6303.0	721.5	386.2	7410.7
May	9274.3	668.9	305.7	10248.9
Jun	8014.1	389.7	301.3	8705.1
Jul	9171.4	245.0	216.2	9632.6
Aug	5857.3	615.4	430.1	6902.8
Sep	6083.8	834.2	321.0	7239.0
Oct	7815.1	601.7	398.2	8815.0
Nov	6161.5	418.8	493.7	7074.0
Total	80295.0	5949.1	4165.6	90409.7

Table III-4: BIONET Prime Time Connect Hours

	BIONET Users except staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	2115.3	399.0	1943.2	4457.5
Jan	1967.8	680.3	1717.9	4366.0
Feb	1982.8	496.6	1488.8	3968.2
Mar	2364.5	633.2	1699.4	4697.1
Apr	2428.6	730.3	1702.7	4861.6
May	3515.5	674.6	1877.6	6067.7
Jun	2533.5	513.8	1642.0	4689.3
Jul	3213.8	679.9	1769.6	5663.3
Aug	2549.4	604.3	1785.3	4939.0
Sep	2442.6	834.5	1668.6	4945.7
Oct	3410.3	1049.5	2003.8	6463.6
Nov	2777.5	828.9	1563.3	5169.7
Total	31301.6	8124.9	20862.0	60288.5

Figure III-4: BIONET's Prime Time Use of the DEC-2060 12/85 - 11/86

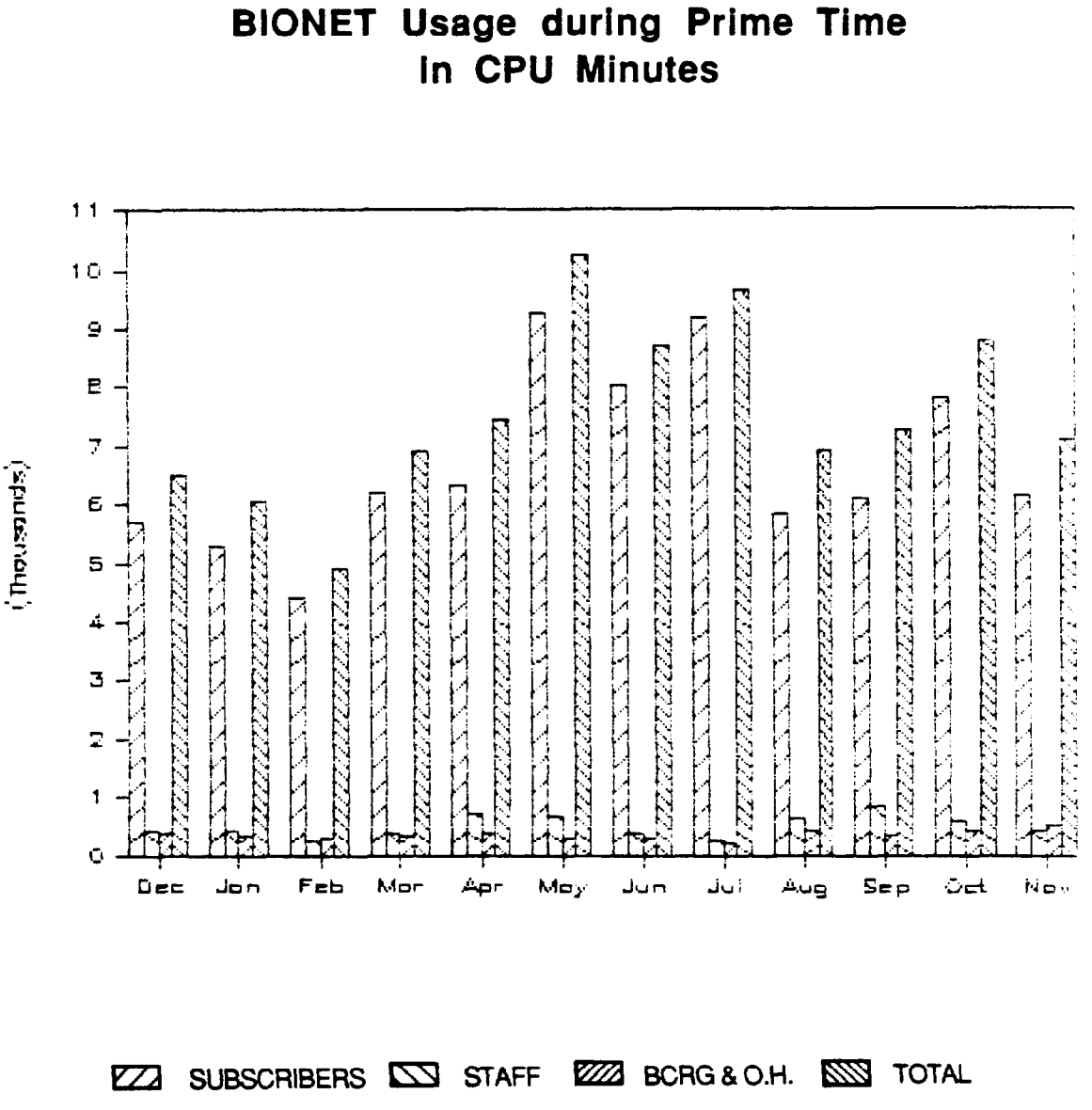


Table III-5: BIONET Non-Prime Time CPU Minutes

	BIONET Users except staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	5888.9	81.7	132.4	6103.0
Jan	5258.4	51.3	940.8	6250.5
Feb	3398.4	31.6	930.5	4360.5
Mar	7836.9	131.2	1007.6	8975.7
Apr	6884.3	173.0	865.2	7922.5
May	15266.7	428.3	962.4	16657.4
Jun	10809.2	60.0	960.6	11829.8
Jul	8905.3	41.5	1038.7	9985.5
Aug	5847.1	46.6	1133.6	7027.3
Sep	7859.8	273.1	905.6	9038.5
Oct	8888.3	302.0	954.0	10144.3
Nov	9877.6	322.6	1107.0	11307.2
Total	96720.9	1942.9	10938.1	109601.9

Table III-6: BIONET Non-Prime Time Connect Hours

	BIONET Users except staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	1311.0	56.3	305.3	1672.6
Jan	1230.1	144.2	2515.0	3889.3
Feb	1317.1	79.8	2362.3	3759.2
Mar	1698.3	113.2	2601.4	4412.9
Apr	1640.7	151.2	2595.7	4387.6
May	2559.7	161.4	2626.9	5348.0
Jun	1769.3	90.6	2362.8	4222.7
Jul	2011.4	196.0	2869.3	5076.7
Aug	1467.9	179.5	2472.9	4120.3
Sep	1416.8	174.3	2736.4	4327.5
Oct	2000.2	155.7	2612.4	4768.3
Nov	1863.0	134.8	2356.7	4354.5
Total	20285.5	1637.0	28416.8	50339.3

Figure III-5: BIONET's Non-Prime Time Use of the DEC-2060, 12/85 - 11/86

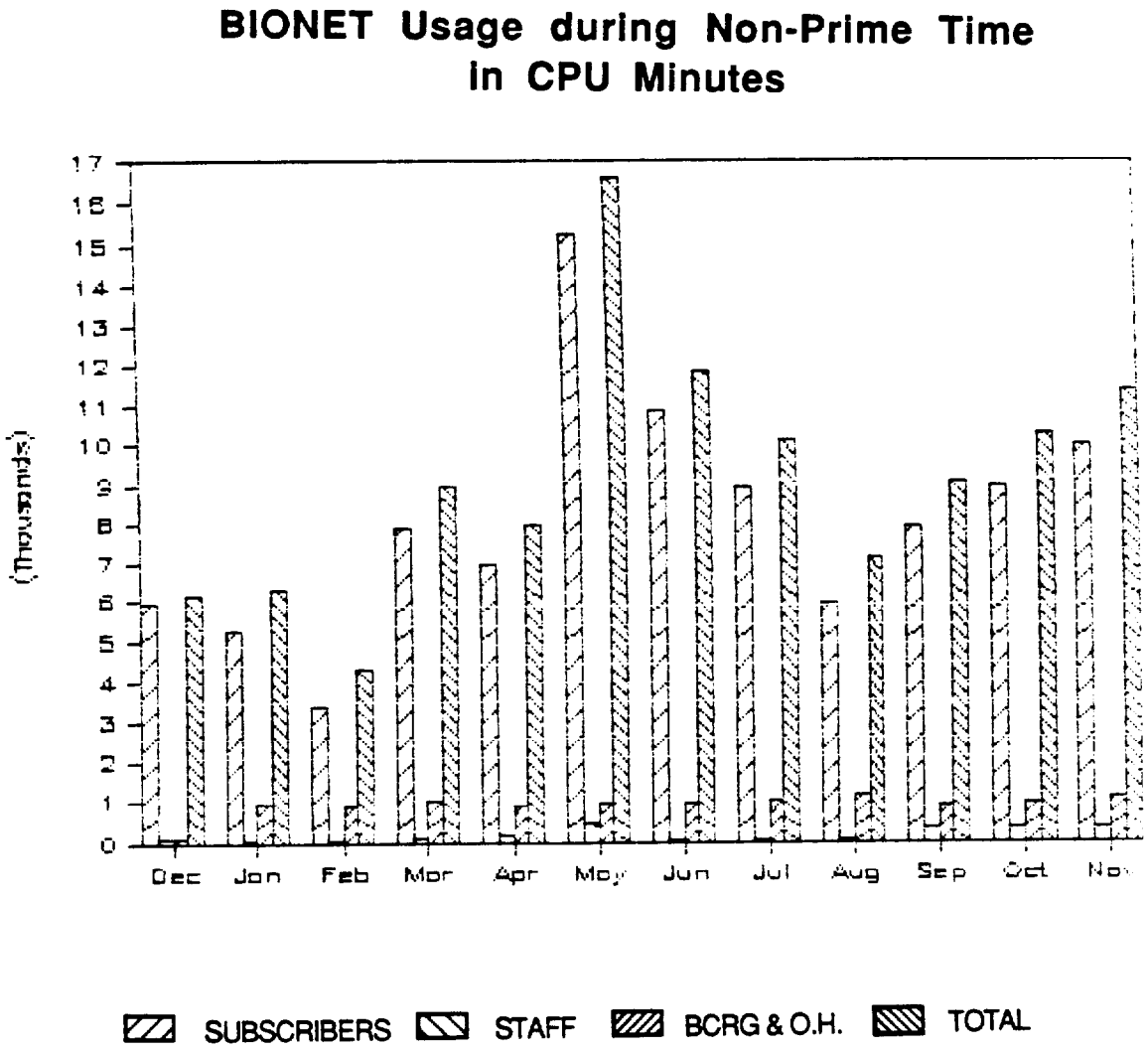


Table III-7: BIONET Total CPU Minutes

	BIONET Users except staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	11621.3	491.2	493.0	12605.5
Jan	10551.8	476.7	1278.5	12307.0
Feb	7805.2	266.4	1198.8	9270.4
Mar	14018.8	515.4	1354.3	15888.5
Apr	13187.3	894.5	1251.4	15333.2
May	24541.0	1097.2	1268.1	26906.3
Jun	18823.3	449.7	1261.9	20534.9
Jul	18076.7	286.5	1254.9	19618.1
Aug	11704.4	662.0	1563.7	13930.1
Sep	13943.6	1107.3	1226.5	16277.4
Oct	16703.4	903.7	1352.2	18959.3
Nov	16039.1	741.4	1600.6	18381.1
Total	177015.9	7892.0	15103.7	200011.6

Table III-8: BIONET Total Connect Hours

	BIONET Users except staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	3426.3	455.3	2248.5	6130.1
Jan	3197.9	824.5	4232.8	8255.2
Feb	3299.9	576.4	3851.1	7727.4
Mar	4062.8	746.4	4300.8	9110.0
Apr	4069.3	881.5	4298.4	9249.2
May	6075.2	836.0	4504.4	11415.6
Jun	4302.8	604.4	4004.8	8912.0
Jul	5225.2	875.9	4638.9	10740.0
Aug	4017.3	783.8	4258.1	9059.2
Sep	3859.4	1008.8	4405.0	9273.2
Oct	5410.5	1205.2	4616.2	11231.9
Nov	4640.5	963.7	3920.0	9524.2
Total	51587.1	9761.9	49278.7	110627.7

Figure III-6: BIONET's Total Use of the DEC-2060, 12/85 - 11/86

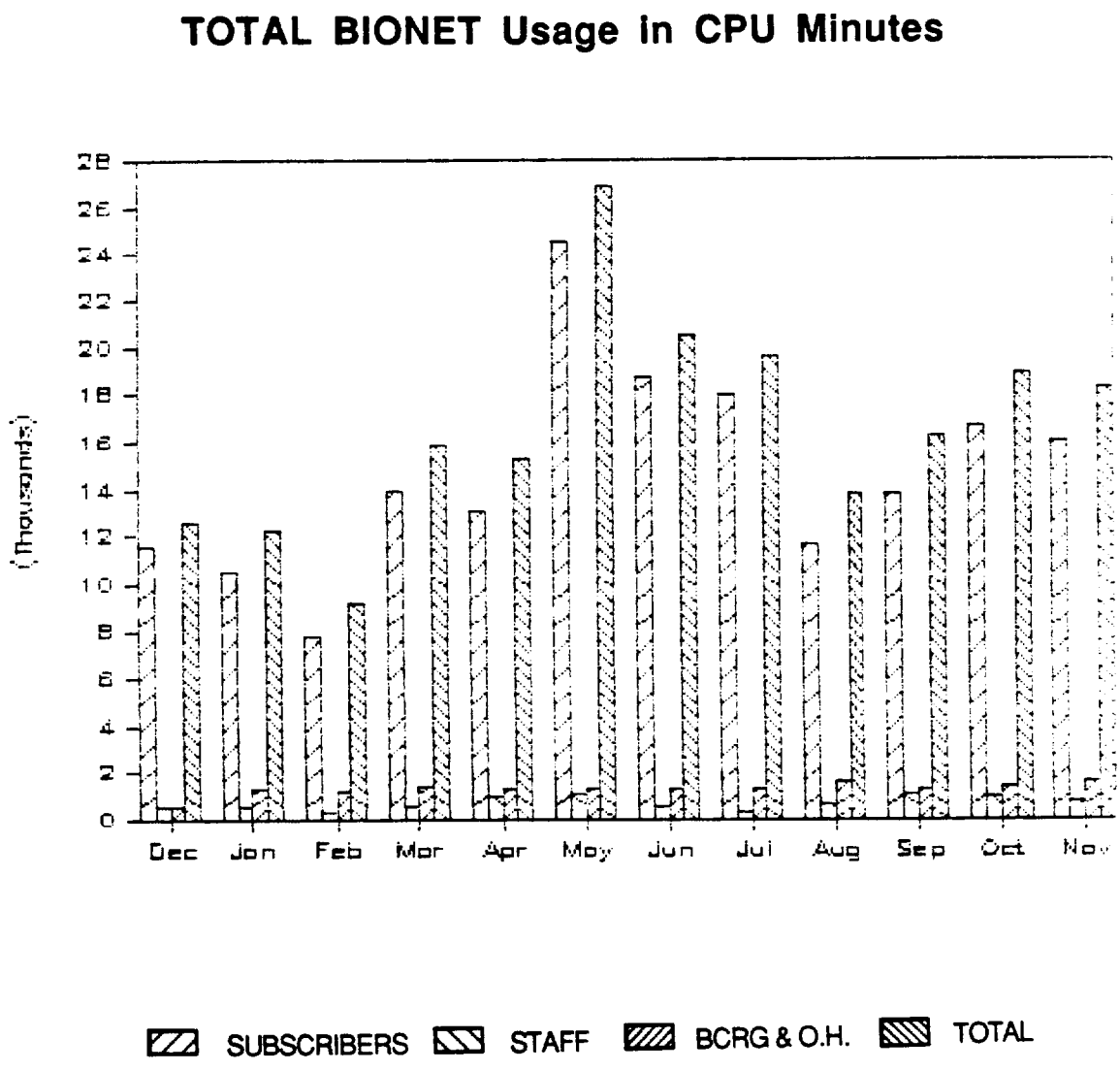
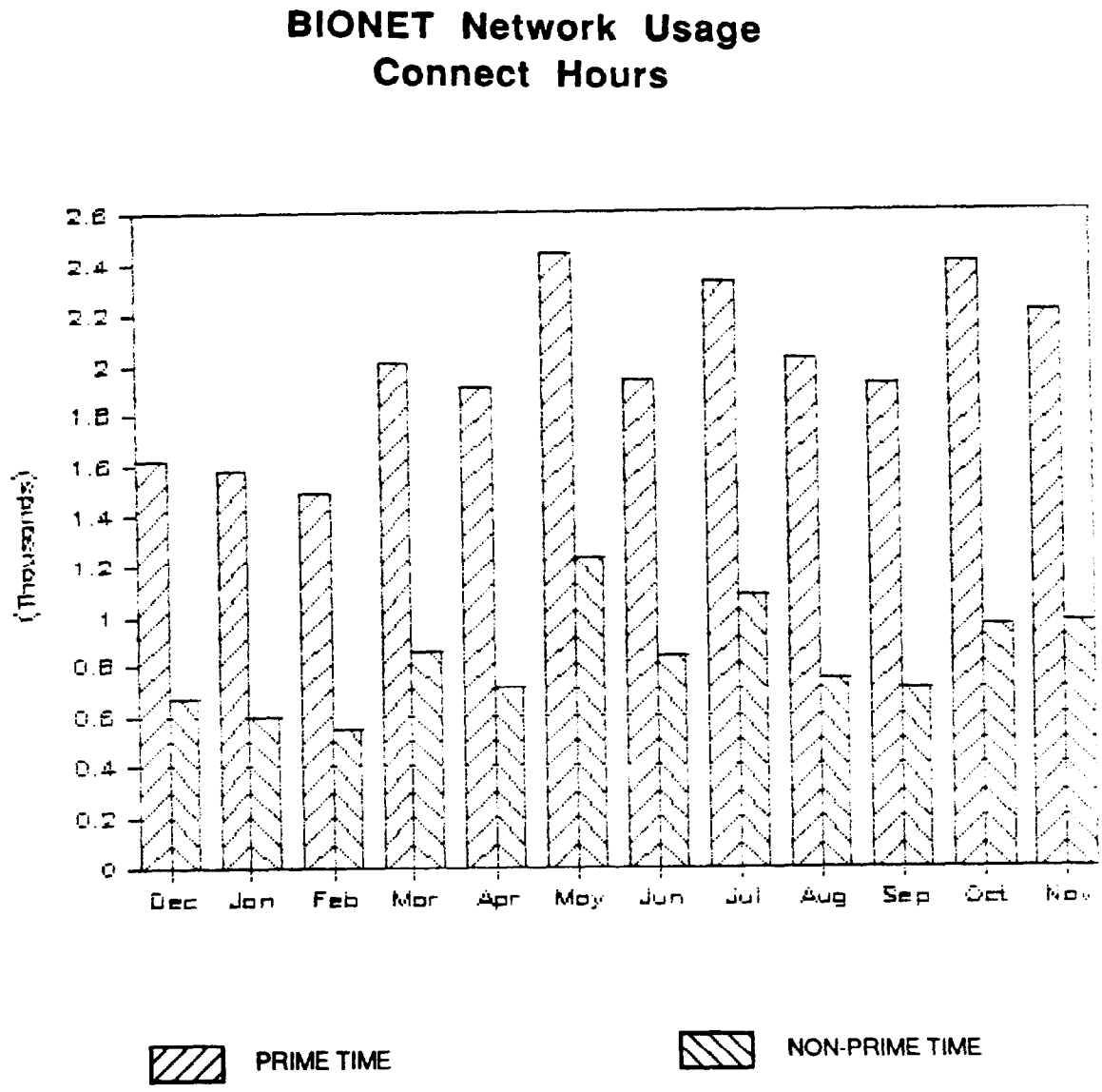


Figure III-7: Total Telenet and UNINET Network Use, 12/85-11/86



interface to the *CLONER* program will be available. This release will be followed shortly by the release of two new modules, *DDMATRIX* for sophisticated dot matrix comparisons of two sequences, and *GENALIGN*, for alignment of multiple nucleic acid or protein sequences.

III.A.5.d. Computer Software - System Library

During the course of the year, the following additions have been made to the system support library described in last year's report.

Operating System - A new release of the TOPS-20 Operating system was installed. This is version 6.1, the version currently supported by Digital Equipment Corp. Several new features are of use to BIONET, including the previously mentioned TCP/IP communication protocol (section 3.A.5.a). Other features include updated utilities, improved filename parsing, and directory access groups which can extend across file systems.

Programming Languages - Several updated versions of the C programming language from Stanford University and SRI International were installed during the course of the year.

System Utilities - A utility TFIND was created to enable users to find their TELENET network access phone number conveniently.

A utility to read Unix TAR format tapes was installed.

A utility MPW was created to suggest memorable, but suitable unique, passwords.

Communication - Several TCP/IP related utilities were installed. FTP permits file transfer under TCP/IP. TELNET permits virtual terminal connections under TCP/IP.

A new version of the BBOARD program permits full integration of electronic bulletin boards with the facilities of the MM message reading program.

III.A.5.e. Computer Software - Contributed Library

Software contributed to BIONET is placed in the <CONTRIBUTED> directory on the DEC-2060, to which only the BIONET community has access. Major software packages produced by BIONET collaborators and implemented on BIONET with the aid of our staff have been summarized under *Collaborative Research*section III.A.2.a, above.

III.A.5.f. Database Library

BIONET provides its users with a large number of different databases in support of molecular biology and molecular genetic research. These range from biological sequence and structure databases through bibliographic and genetic map databases. This year two different DNA sequence databases, the EMBL and NIH Genbank databases have grown in parallel and the overlap in their content has increased as they approach a more common format. Figures III-8 and III-9 show the growth in these databases with each release. In 1985 we maintained one release of these databases on BIONET for each release from the source. During the past year we have made one release every two months. In addition we have prepared special update files that allow the users to screen or search just the new sequences in each update. This has been a considerable simplification for the BIONET users and has helped save on CPU time as well. Our database releases have usually occured 2-3 weeks after obtaining the tapes from NIH GenBank or the EMBL.

The DNA sequence databases are used by BIONET scientists both as a source of sequence data and for homology searches. Two major types of searches are performed. Those involving the QUEST program search the DNA databases for interesting consensus sequences of functional importance. The second type of database search looks for similarities or homologies with sequences using the IFIND program. It is hard to estimate the total number of such searches since most users search different parts of the database rather than the entire database. We do know that the most frequently searched portions of the database are the files containing human DNA sequences. They are searched an average of 166 times per month (5-6 times per day). The rate of use of other sections of the database vary between 20 and 60 times per month. Perhaps a better measure of the use of the databases comes from the number of times that the QUEST and IFIND programs are used on the system. The QUEST program is used 864 times per month to search for consensus sequences and IFIND is used 1115 times per month to search for homologies. This amounts to nearly 2000 searches utilizing the databases each month.

The EMBL database is used considerably less than the NIH Genbank database with its files being referenced only 27 times per month. We attribute this to the great amount of overlap between the NIH GenBank and the EMBL databases as well as to the more frequent releases from Genbank. As these two resources work more closely with each other, BIONET may chose to release only the NIH Genbank database.

Protein Sequence Database - BIONET makes the Protein Identification Resources protein sequence collection available for similar homology searching. This database is searched 468 times per month for homology using the IFIND program and the same database is searched 60 times per month using Bill Pearson's XFASTP program. This makes protein database searching the second most highly used database on the BIONET resource. The growth of this database and the number of releases are shown in Figure III-10.